心理科学进展 2023, Vol. 31, No. 1, 60–77
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2023.00060

·研究前沿(Regular Articles)·

算法决策趋避的过程动机理论*

谢才凤[1] 邬家骅[1] 许丽颖[2] 喻 丰[1] 张语嫣[1] 谢莹莹[3]

([1]武汉大学哲学学院心理学系, 武汉 430072) ([2]清华大学马克思主义学院, 北京 100084)
([3]灵山县那隆镇中心校, 广西 钦州 535414)

摘 要 算法常用于决策, 但相较于由人类做出的决策, 即便内容相同, 算法决策更容易引起个体反应的分化, 此即算法决策趋避。趋近指个体认为算法的决策比人类的更加公平、含有更少的偏见和歧视、也更能信任和接受, 回避则与之相反。算法决策趋避的过程动机理论用以解释趋避现象, 归纳了人与算法交互所经历的原初行为互动、建立类社会关系和形成身份认同三个阶段, 阐述了各阶段中认知、关系和存在三种动机引发个体的趋避反应。未来研究可着眼于人性化知觉、群际感知对算法决策趋避的影响, 并以更社会性的视角来探究算法决策趋避的逆转过程和其他可能的心理动机。

关键词 算法决策, 人类决策, 决策趋避, 心理动机, 人–算法交互
分类号 B842

1 算法决策或趋或避

算法常用于提供建议、判断和预测, 因之产生算法决策(Burton et al., 2020; Silva & Kenney, 2018)。其为增强决策、辅助决策、专家系统和诊断辅助等概念的总称, 主要载体包括机器(人)、自动化和人工智能(Bigman & Gray, 2018; Malle et al., 2015; Araujo et al., 2020)。旧有研究对算法决策主体未加细致区分(Lee, 2018; Möhlmann & Zalmanson, 2017), 其可被视为以软件、计算机、自动化系统、机器(人)或其它人工智能体的形式基于决策规则或统计模型所执行的判断、预测、建议、辅助决策、决策等功能。通常决策主体为人类, 但人类决策囿于客观信息不足与主观偏见(Gilovich et al., 2002), 不及算法决策客观(Lindebaum et al., 2020)、精确(Donnelly, 2017)、迅速(Bonnefon et al., 2016)与低廉(Esteva et al., 2017)。算法决策因之应用于医疗(Biró et al., 2021)、司法(Grgić-Hlača et al., 2019)、经济(Harvey et al., 2017)、交通(Badue et al., 2020)和招聘(Cheng & Hackett, 2021; Raisch & Krakoswki, 2021)等场景。当算法渐入生活, 民众将如何看待?

对决策者的反应取决于形态感知与关系类型。算法基于无形代码, 虽非人类, 但通过拟人化感知可诱发主体知识(elicited agent knowledge), 使人视之为社会成员(喻丰, 许丽颖, 2020)。拟人化是将人类独有的特征、动机、意向或心理状态赋予非人对象的过程(许丽颖 等, 2017; Epley et al., 2007)。人们可以通过将姓名、性别、肢体等人类属性赋予自动驾驶汽车和机器人(Malle et al., 2016; Waytz et al., 2014), 亦可将自由意志、能动性和体验性等心智归因于算法而将其拟人化(许丽颖 等, 2022; Bigman & Gray, 2018; van der Woerdt & Haselager, 2019)。拟人化之算法在社会认知存在链(social cognitive chain of being)上位置升高(Brandt & Reyna, 2011), 然仍与人相去甚远。原住民将外来者视为威胁而抗拒, 白种人将黑种人当作奴仆而欺压, 其实质是群体间的非人化(dehumanization; Cooley & Payne, 2017, 2019; Hehman et al., 2018; Stewart & Morris, 2021)。算法同样被感知为缺乏完整心智(Bigman & Gray, 2018), 或被视为“人类生存最大的威胁”

收稿日期: 2022-02-19
* 国家社科基金青年项目(20CZX059); 国家自然科学基金青年项目(72101132)。
通信作者: 许丽颖, E-mail: liyingxu@mail.tsinghua.edu.cn
喻丰, E-mail: psychpedia@whu.edu.cn

(Bostrom, 2014; McFarland, 2014), 此亦非人化。接受新异之物而将之视作人之过程与人类认知外群体可进行比较。

人类对待外群体的态度以及非人化过程受意识形态影响(Jost et al., 2003; Jost et al., 2019)。保守主义指倾向于选择安全、传统和常规形式的制度和行为(Wilson, 1973), 与保持传统、追求稳定、抵制变革的偏好相关(Graham et al., 2009; Jost et al., 2003)。而自由主义则与追求创新、倡导变革和接受平等的偏好有关(Jost et al., 2008a)。自由主义者主张优待弱势群体, 对移民等社会地位较低的外群体持更积极的态度和更高的接受度(趋近), 而保守主义者则有更强烈的内群体偏好, 对穆斯林、外国人和非法移民等外群体的敌意和歧视也更强烈(回避) (e.g., Iacoviello & Spears, 2021; Jost et al., 2003; Kugler et al., 2014; Stewart et al., 2019)。有证据表明, 当听闻黑人被警察枪杀时, 自由主义者更能将此事与种族主义联系起来(Cooley et al., 2019)。而面对移民, 保守主义者更倾向把外群体当作威胁, 从而表现出更强烈的偏见(Stewart & Morris, 2021)。

既然意识形态区分了人类于心理和行为层面对外群体的趋避反应, 那么这种社会认知在链上的由高视低亦将算法包括在内(Grgić-Hlača, Zafar, et al., 2018; Wetherell et al., 2013; Stewart et al., 2019)。算法(或机器人)因其自主性构成潜在的风险 (Bostrom, 2014) 与威胁 (McFarland, 2014; McClure, 2018)。而当威胁信号出现, 人便会有“战或逃”反应(fight or flight response), 并体现在生理唤醒与防御行为中(Blanchard et al., 2001; Suresh et al., 2014)。研究显示, 人在感知到机器人或自动驾驶汽车即将危及自身时会出现踢打、轻拍或躲避等战或逃反应(Mahadevan et al., 2018)。然而, 战或逃反应与算法趋避不尽相同。不同之处在于, 算法决策能造成的威胁类型更为多样(Lasota et al., 2017), 人除了生理与无意识行为之外, 还会产生认知、情感方面的反应(Mahadevan et al., 2018)。相同处则体现在二者均反映对刺激来源趋近或回避的态度。

进化观认为物种若不能辨别环境刺激的利弊, 便会遭到淘汰, 因此趋避是有机体对环境刺激最基本的反应(Schneirla, 1959; Zajonc, 1998)。被认为是积极的、期待的事件会激发人的趋近行为, 反之则引起回避行为(Bargh et al., 1996; Elliot, 1999)。以外群体导致的趋避反应为例: 黑人因表现出与消极刻板印象(如学业不良)相反的特征时, 会被指控为饰伪而成为威胁信号(Neal-Barnett et al., 2010), 白人的回避便随之出现(Cooley & Payne, 2017, 2019; Hehman et al., 2018; Stewart & Morris, 2021); 相反, 人们也会在意识到同性恋群体不构成威胁后表现出趋近的态度(Turner et al., 2013)。算法决策是更为复杂的刺激源, 对其风险和利弊的判断决定人们的趋避反应。一方面, 现有的算法决策安全、高效、低廉的优点能使民众产生趋近反应, 并表现为信任与接纳(Parasuraman & Riley, 1997)。另一方面, 因技术所限, 算法决策仍存在风险, 甚至引起失业危及社会导致回避反应, 表现为质疑或抵制(Nolan et al., 2016)。

另有计算机作为社会参与者理论(computers are social actors theory)认为, 人们会以对待社会成员的方式对待计算机, 将规范、类属和期望代入人机交互过程, 并表现出与人际交互相似的社交反应和社会联系(Nass et al., 1997; Nass & Moon, 2000)。据此似乎可以推测, 鉴于算法与人类思维的相似性, 民众仅根据决策内容做出趋避反应。但与上述推断不符的是, 多数研究发现, 即便内容相同的决策经由不同主体(人类或算法), 受众在对决策主体的公平感、信任度和道德责备方面表现出明显的差异——对算法决策无论趋避均有更强烈的反应倾向(Langer et al., 2021; Ötting & Maier, 2018; Suen et al., 2019; Scheutz & Malle, 2021)。

因现有研究无法概括这一不对称现象, 本文称之为算法决策趋避, 即相较于人类决策, 人们对算法做出同样的决策产生更明显的趋近或回避反应。具体而言, 算法决策趋近指经比较后认为算法决策含有更少的偏见和歧视、更为公平, 也更容易产生信任和接受的态度; 算法决策回避则反之。需说明的是, 定义中含有对人类和算法的比较, 而非仅对算法单一主体决策的趋近或回避反应。且因研究采用不同计分方式而无法使用统一标准来明确和衡量算法单一决策趋避的具体程度大小(e.g., Acikgoz et al., 2020; Bigman et al., 2020; Voiklis et al., 2016), 只能暂且将人们对算法决策和人类决策的反应进行比较而得出趋避倾向。

一些情况下, 相比于人类决策人们更趋近算法决策。如在面试推荐(Jago & Laurin, 2021)、申请休

假(Schlicker et al., 2021)、大学录取(Marcinkowski et al., 2020)、司法或健康决策(Araujo et al., 2020)以及撰写或审核新闻(Waddell, 2019; Wang, 2021)等情境下，多数人认为算法决策比人类决策具有更少的偏见和歧视，更多的公平。当意识到面试官在招聘中存在歧视的可能后，人们更倾向于接受来自算法的评估，也更容易忽略算法决策所包含的歧视(Bonezzi & Ostinelli, 2021)。算法决策产生的歧视较之人类决策引发了更少的道德愤慨(Bigman et al., 2020)。人们也更信任机器审核的新闻(Wang, 2021)，认为算法撰写的新闻比人类撰写的更专业和可信(Graefe et al., 2018)，尤其是体育新闻(Wölker & Powell, 2021)或需要大量信息处理的新闻(Liu & Wei, 2019)。在涉及对个人身体状况、吸引力、奖金评估的情境中，用户对算法的接受程度更高(Logg et al., 2019)，即便将算法和人类的建议相结合也难以改变这种单纯对算法的偏好(Logg, 2017; Logg et al., 2019)。

在另一些情况下，人们则会产生更为回避的反应。例如人们不放心将关系重大的决策委托给人工智能(Leyer & Schneider, 2019)。在涉及人力资源管理的面试、裁员、升职和绩效审查等环节中，员工认为算法决策更不公平(Acikgoz et al., 2020; Diab et al., 2011; Newman et al., 2020; Nørskov et al., 2020)。在医疗健康领域，患者更信任人类医生而非医疗算法(Promberger & Baron, 2006)，难以接受来自人工智能的医疗服务和建议，为人工智能支付医疗费用的意愿也更弱(Longoni et al., 2019)。在金融投资领域，人们更信任人类的预测而非算法统计的结果(Önkal et al., 2009)，以消极行为反抗算法决策(Filiz et al., 2021)，或是在算法出错后坚信自己的判断(Dietvorst et al., 2015)。在司法领域，算法提供的错误信息使当事人在败诉后倾向采取羞辱、报复和抗议手段表达不满，并造成更严重的负面情感反应(Ireland, 2020)，如失望、愤怒和沮丧等(Lee, 2018)。

算法决策趋避在道德场景中表现得更为复杂。人们通常认为算法不适合做出道德决策，即使人类和算法的能力无异(Jago, 2019)。处在两难困境下，机器只被允许做出牺牲少数人以拯救多数人的决策，若不如此，它会受到更多的道德错误指责与责备(Voiklis et al., 2016)。相比之下，人类做出不作为和义务论的决定将受到比机器人更少的道德责备，但当选择作为和功利论时，人类受到的道德责备反而与机器人无异或比机器人更多(Malle et al., 2015; Scheutz & Malle, 2021)。当自动驾驶汽车和人类司机因同样的过错而造成无法避免的事故时，人们给自动驾驶汽车分配更少的责任，将更多责任归因于制造商和政府(Li et al., 2016)。在由人和机器共同控制的情况下，机器驾驶员对错误决策所负的责任也始终都低于人类(Awad et al., 2020)。

2 算法决策趋避之源

人们为何表现出算法决策趋避？类比人对外群体的态度，意识形态可能是算法决策趋避的影响因素。具体而言，自由主义与算法决策趋近相关，而保守主义则与算法决策回避有关。

民众的保守主义或自由主义倾向主要取决于相互关联的两个方面：倡导还是抵制社会变革，以及拒绝还是接受不平等(Jost et al., 2003; Jost et al., 2009)。自由主义者对不平等制度更敏感和反感(Jost et al., 2009; Napier & Jost, 2008)，也更支持社会变革(e.g., Anderson & Singer, 2008; Jost et al., 2008a)。保守主义者则有更强的系统合理化信念(Jost et al., 2008b)，会为了降低威胁和不确定性而抵制变革和接受不平等(Graham et al., 2009; Jost et al., 2003)。

自由主义者更相信算法决策(Gauchat, 2012)，更支持推广无人驾驶汽车以服务于儿童、老年人和残疾人(Dixon et al., 2020)，也更赞成立法保护人工智能和机器人(Lima et al., 2020; Martínez & Winter, 2021)。另有研究表明，在控制机器启发式先验信念后，自由主义者认为由算法撰写的反对特朗普的新闻具有更少的偏见(Jia & Liu, 2021)。保守主义者则常将机器人视为威胁(Oleksy & Wnuk, 2021)。由于机器人介入医疗或法律领域会对人类独特性(human uniqueness)构成威胁，保守主义者持更消极的评价与看法(Han et al., 2021)。保守主义者也会认为医疗人工智能和自动驾驶汽车有更大的风险而不予信任、拒绝使用，并支持出台限制人工智能的政策(Castelo & Ward, 2021; Peng, 2020)。保守主义者也认为由算法撰写的新闻具有更多的偏见(Jia & Liu, 2021; Waddell, 2019)。而意识形态居于中间者则没有对决策者明显的趋避倾向(Waddell, 2019)。

算法决策趋避与意识形态有关，而意识形态又受动机性因素影响。动机社会认知理论(Motivated Social Cognition Model)认为，意识形态由三类动机塑造，即认知动机、关系动机和存在动机(Jost et al., 2003; Jost et al., 2008a; Jost et al., 2009; Jost & Amodio, 2012)。该理论解释了人们如何在不同情境中选择自身的意识形态。比如经历过 9·11 恐怖袭击事件的人会出于存在动机而趋向于支持保守主义(Bonanno & Jost, 2006)。面对算法这样的新兴决策主体，人们也会出于认知、关系和存在动机而形成态度及信念，既而选择趋避。

人机关系类型也会影响算法决策趋避。类社会互动理论(parasocial interaction theory)认为，与虚拟人物的互动，会产生虚幻的社会关系(Hartmann, 2008, Horton & Wohl, 1956; Stern et al., 2007)。人们会以处理人际互动相似的方式去处理类社会互动中的问题，两类关系是平行发展的 (Horton & Wohl, 1956; R. B. Rubin & A. M. Rubin, 2001)。比如顾客与购物网站、语音助手及推荐系统进行类社会互动并建立类社会关系，并接受对方建议(Chung & Cho, 2017; Tran et al., 2019; Whang & Im, 2021)。

人际关系中的吸引、相似性感知和移情等因素同样作用于类社会关系，如长期且深入的接触能加强对虚拟人物的亲密感知(Davis, 1973; Horton & Whole, 1956; R. B. Rubin & A. M. Rubin, 2001)。当人们不满足于虚幻的关系时，便试图通过实际接触来建立更为真实的联系(Horton & Whole, 1956)。因此人们与计算机之间会产生强烈的相似性感知(Nass et al., 1996)，而感知相似性有助于建立心理联结(Amiot et al., 2020)，并形成社会身份认同(van Vugt & Hart, 2004)。认同的过程是动态的，其产生于人际交互中，由与其他人、群体或文明的关系来界定，反映了自我与他人、群体或文明之间的关系(Tajfel, 1974; Tajfel, 2010)。由此可推断，在人与算法的类社会互动中，民众大致会经历三个阶段，即原初行为互动阶段、建立类社会关系阶段和形成身份认同阶段。

人与算法交互的三个阶段是依次递进的，人们会出于认知、关系和存在动机需要来应对风险和权衡利弊，最后选择趋近或回避算法。在原初行为互动阶段，人们会感知算法决策存在并与之接触，从而产生减少不确定性、复杂性和模糊性的认知动机需要(Jost et al., 2009)，机器启发式、对算法决策的准确性和客观性认知以及熟悉度可以满足这种需要，进而使人们趋近算法决策，反之则回避。在互动过程中，人们也会渴望与算法建立“人际关系”而产生关系动机(Jost et al., 2009)。在建立类社会关系过程中，算法心智缺乏意味着其道德地位和道德能力的缺失，使用不仅存在风险还减少将决策责任或责备转移给算法的可能。而且算法决策虽能减少偏见和不公平等(Bigman et al., 2021; Howard et al., 2020)，但也导致了人际接触的减少，于人们兼具利弊，使得人们表现出算法决策趋避。随着类社会关系的建立，人与算法产生心理联结进而形成身份认同。然而，算法对人类构成了现实威胁和身份威胁(Huang et al., 2021; Yogeeswaran et al., 2016)。出于应对威胁的需要(Jost et al., 2009)，存在动机也影响对算法决策的趋避。基于以上，构建出算法决策趋避的过程动机理论，见图 1。

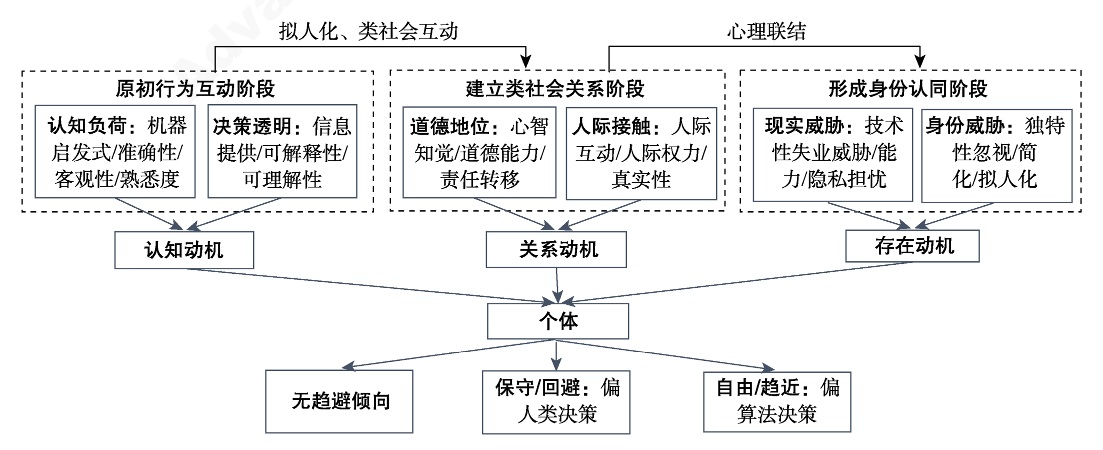


图 1　算法决策趋避的过程动机理论框架

3 算法决策趋避过程

3.1 原初行为互动阶段

在与算法互动之初，人们会因信息缺乏而认为算法决策是高深莫测的(Yeomans et al., 2019)，出于减少不确定性、复杂性或模糊性的需要而感知风险和权衡利弊，进而产生算法决策趋避。

3.1.1 认知负荷

机器启发式(machine heuristic)以节省认知资源与降低认知负荷的方式处理信息，并降低不确定性，从而产生算法决策趋近(Fiske & Taylor, 1991; Gary & Wood, 2011; Todd & Benbasat, 1994)。机器启发式指人在接触机器后会自动启动刻板印象，认为机器没有感觉、思想和情感，更加客观中立，能以更精确安全的方式处理信息和执行任务，从而对机器产生积极反应(Sundar, 2008; Sundar & Kim, 2019)。形态–媒介–交互性–适航性(Modality-Agency-Interactivity-Navigability)模型表明，媒介可以触发认知启发式，帮助人们对信息来源及其内容做出可信度判断(Sundar, 2008)。因此当机器作为信息媒介，便能触发机器启发式，将算法决策知觉为客观且无偏见的，进而认为算法决策比更加公平可信(Grgić-Hlača, Redmiles, et al., 2018; Helberger et al., 2020; Wang, 2021)。

启发式也可能先于交互，以认知图式出现(Sundar et al., 2020)。人们通常认为算法决策较之人类决策快速、一致、准确，而更加客观公正(Haenssle et al., 2018; Jago & Laurin, 2021)。认为算法建议更专业有效的信念也会增加信任进而提高算法决策的利用度(Kramer et al., 2018)。在新闻和交通等领域中，算法决策通常因无主观意图和偏好、决策过程标准化的印象被认为中立客观，人们对其公平性感知和信任度因此提升(Howard et al., 2020; Miller & Keiser, 2021; Tandoc et al., 2020)。但对人类是实用、合法、全面、专业和一致的启发式加工，也能使由人主导的面试获取更高的接受度(Diab et al., 2011)。

对算法的熟悉度可以激活启发式认知，从而降低认知负荷(Kahneman, 2003)。比如熟悉导航网站的用户信息处理的速度更快，关联的认知负荷也更低(Hu et al., 2017)。在熟悉算法并习惯由算法主导的特定决策之后(Kramer et al., 2018)，人们会一改最初对新技术的排斥态度，转而趋近算法决策(Parasuraman & Riley, 1997)，如更能接受机器人的道德决策(Komatsu, 2016)和司法裁决(Ireland, 2020)。

3.1.2 决策透明

人对决策的反应也取决于决策者所提供的信息以及是否加以解释(Dodge et al., 2019)，因此透明性也是算法决策趋避的原因，即可解释性和可理解性，指人们对算法决策过程和结果相关信息的可获取程度(Shin & Park, 2019; Shin, 2020)。算法不会呈现足够的信息并解释的特点会引起人的不安与回避(Acikgoz et al., 2020; Langer et al., 2018)。倘若完全缺乏关于决策的解释，人们认为自动化决策比人类决策的信息公平和程序公平感更强(Schlicker et al., 2021)。但提供解释会让人们觉得自动化决策比人类决策更公平(Schoeffer et al., 2021)，继而减少对算法的拒绝(Yeomans et al., 2019)。对决策结果进行解释也会产生减轻算法责任或增强对算法决策公平感知的极端后果(Lee et al., 2019)。研究表明，呈现数据会给人以公平的感知，隐瞒或不对数据加以解释则会降低决策的透明性，破坏原本公平的印象(Dodge et al., 2019)。

算法决策中的可解释性也指算法应用中的方法和技术被人们理解的程度(Ehsan & Riedl, 2019)，人们通常很难理解算法或推荐系统是如何运作的(Kroll et al., 2017)，也因对算法缺乏了解而产生不信任，进而拒绝使用算法(Logg et al., 2019; Prahl & van Swol, 2017)。Yeomans 等人(2019)指出推荐系统仅仅做到准确是不够的，它们还必须被理解。相比之下，医生经与患者沟通后能使其理解决策，而这也造成患者抵制算法诊断皮肤癌的结果(Cadario et al., 2021)。少量研究也表明，当人们了解算法的工作原理并意识到其中潜在的偏见后，原先的算法偏好就会消失，即对算法学习的担忧会破坏算法没有偏见的固有启发式，从而引起算法决策回避(Jago & Laurin, 2021)。

3.2 建立类社会关系阶段

在与算法实际互动后，人们可能出于满足归属感和社会认同的需要而产生关系动机，并与之建立类社会关系(Han & Yang, 2018)。而在建构关系的过程中，算法道德地位以及人际接触影响了算法决策趋避。

3.2.1 道德地位

算法通常由于缺乏心智而被认为不具备完全

的道德地位(Bryson, 2020)。道德的本质是心智知觉，即道德被认为需要完整的人类心智，人们通过目标的心智水平来判断其是否有完全的道德地位(Bastian et al., 2012; Gray et al., 2012)。心智知觉理论(mind perception theory)认为，心智通过两个维度被感知，即能动性(agency)和体验性(experience; Gray et al., 2007)。能动性包括思考、计划、记忆、行动等能力，体验性则对应饥饿、恐惧、愉悦、欲求等能力(Gray et al., 2012)。机器具有一定的思考、计划、记忆和自控的能力，比如它们可以进行大量的复杂计算、与人沟通或下棋等(Silver et al., 2017)，因此具有一定的能动性(Gray et al., 2007; Gray & Wegner, 2012)。但是与人类相比，机器因为不具体验性而缺乏完整的心智(Brink et al., 2019; Reinecke et al., 2021; Swiderska & Kuster, 2020)，因此机器被认为是有限道德主体，道德地位低于人类。具体而言，由于缺乏自主性、道德推理、沟通与判断行为后果等能力(Cushman, 2008; Malle, 2016)，机器被认为不具备做出道德决策的能力。

由于无法感知算法的完整心智和道德能力，人们产生算法决策趋避。一方面，人们更喜欢由人类而非机器做出关乎生死的道德决策，即使意识到机器专业性显著优于人类，或机器决策可以带来积极的结果，这种算法拒绝仍然存在(Bigman & Gray, 2018)。另一方面，对算法缺乏情感和体验的认知反而使得人们认为算法决策具有更少的歧视，进而认为算法决策更公平(Helberger et al., 2020; Jago & Laurin, 2021; Noble et al., 2021)。心智缺乏决定了算法仅具有一定的道德地位，这让人们觉得算法的道德能力有所缺失。也正因如此，在与道德相关的领域中，人们更可能将影响第三方的决策任务委托给人类而不是机器，并给予人类的决策以更高的评价(Gogoll & Uhl, 2018)，在涉及道德的投资中也是如此(Niszczota & Kaszás, 2020)。

道德地位的缺失也意味算法不能作为完整的责任主体，这会减少人类将决策责任或责备转移给算法的可能，进而使得人们可能回避算法决策。人类有能力承担决策责任或负最终决策责任，而算法却缺乏承担责任的相应能力(Promberger & Baron, 2006)。对责任分配的考量也会影响人们对决策主体的选择(Steffel et al., 2016; Steffel & Williams, 2018)。作为决策主体，算法不能承担决策失误的责任而人类却可以，出于将部分责任转移至他人而使得自己无需承担全部责任的考虑，人们会出现算法回避反应(Bonaccio & Dalal, 2006; Promberger & Baron, 2006)。然而，人们也可能故意使用诊断辅助或以计算机遴选员工的方法逃避责任，因为在决策产生负面结果后，他们将受到了更少的批评和指责(M. V. Pezzo & S. P. Pezzo, 2006; Nolan et al., 2016)。责任转移显然是算法决策趋避的影响因素之一，但会因具体情境而不同，未来可以就此做更深入的探讨。

3.2.2　人际接触

由于人与算法之差异，人们可能会在人机交互中感受到更少的接触而导致算法决策回避。与人类相比，算法因决策过程缺乏协商沟通，致使人们感受到更多的不平等(Acikgoz et al., 2020; Glikson & Woolley, 2020; Helberger et al., 2020; Noble et al., 2021)。例如，面试者被算法遴选后，会报告自己的人性化下降、缺乏双向沟通、没有表现和复议的机会及更差的待遇，进而认为算法决策不及人类公平(Acikgoz et al., 2020; Kaibel et al., 2019; Noble et al., 2021)。自动化面试也会减少沟通，降低应聘者的存在感，造成不接受的态度(Langer et al., 2019)。

交互双方的特征也会通过影响人们在决策过程中的互动性感知，进而导致算法决策趋避。研究表明，在人际互动中有影响力者会认为人类的调解比算法更公平 (Lee & Baykal, 2017)。人们会觉得算法决策缺少温情、善意和对使用者的尊重，进而认为算法决策不如人类决策公平(Wang, 2018; Kaibel et al., 2019; Langer et al., 2021)。Jago (2019)的研究表明，人们觉得算法在音乐与绘画创作中的表现不如人类工作真实，这种真实性的差异使他们认为算法做出的道德决策比之人类更不道德，也更喜欢人类决策。与算法交互中产生的不适感也会造成算法决策回避，但决策的效用也能在一定程度上抑制回避(Castelo et al., 2019)。遭遇歧视的经历，也会使个体认为算法的筛选和选拔过程更为合理，给予自己更多表现的机会，并认为所要加入组织更具吸引力(Kaibel et al., 2019)。

3.3　形成身份认同阶段

群际关系研究表明，人们将人类视为内群体，将动植物或算法等视为外群体(Turner et al., 1987)，并倾向视外群体为威胁的来源。当机器人的现实

威胁和身份威胁凸显，出于应对威胁、寻求安全感的存在动机需要，人们可能对算法决策产生相应的趋避反应(Huang et al., 2021; Yogeeswaran et al., 2016; Złotowski et al., 2017)。

3.3.1 现实威胁

现实威胁涉及对内群体的资源、工作或安全的威胁，也是群际偏见的预测因素(Riek et al., 2006)。因此，将机器人视为现实威胁会产生算法决策回避。

使用新技术可能会带来失业等社会问题(Benzell et al., 2015)，这会迫使民众回避算法决策。对技术性失业(technological unemployment)的担忧会放大算法的威胁(Headrick, 2009; McClure, 2018; Radinsky, 2015)。例如职员会在算法主导的评价体系下难以感知自身价值，因而抵制计算机的标准化决策(Meehl, 1986)。在 Nolan 等人的研究中发现，虽然使用计算机程序进行雇佣决策可以使招聘被认为是稳定且可靠的，但关于雇佣者认为雇员对雇佣决策过程的因果关系(或掌控感)更少的担忧，会使得雇员自己对雇佣决策感知价值(即对技术性失业的恐惧)的关注增加，随后降低雇员使用这些标准化决策的意图(Nolan et al., 2016; Nolan et al., 2020)。

机器人的能力也会引起人的警惕，进而导致算法决策趋避。研究表明，机器人的能力越强，在任务中的表现越好，对人类构成的现实威胁就越大(Yogeeswaran et al., 2016)，人们的信任程度也越低(Hancock et al., 2011)。比如，在做出关乎生命安全的手术决策中，人们几乎宁愿选择一个普通的医生而非专家机器(Bigman & Gray, 2018)。然而机器的能力并非仅是构成威胁，也可能造成偏好。有证据显示，普通民众更重视、信任和接受来自算法而非普通人的建议(Logg et al., 2019; Madhavan & Wiegmann, 2007)，未来研究应进一步探讨算法能力影响算法决策反应的边界条件。

当算法在交互过程中掌握大量的个人隐私后，可能会成为一种现实威胁。譬如，购物和浏览新闻的过程需要依靠算法进行数据收集 (Kozyreva et al., 2021)，出于对隐私泄露的担心，以及害怕被持续监控或成为个性化营销牺牲品的担忧，人们会回避算法决策(Auxier et al., 2019; Krupp et al., 2017)。个体对隐私泄露的担忧越高，其越怀疑算法决策的道德性进而认为算法决策更不公平，但当人们相信自己可以保护自身的在线隐私或对在线信息拥有更多控制权时，就更有可能认为自动化决策是公平且有用的(Araujo et al., 2020)。此外，坚信机器在处理隐私信息方面比人类更安全、更值得信任的消费者在预订机票时也更愿意授权个人信息(Sundar & Kim, 2019)。

3.3.2 身份威胁

身份威胁指对自身特殊性、价值观和独特性被忽视的担忧(Riek et al., 2006)。高度拟人化的机器人融入社会后，人与机器的区别便会模糊，并对人类身份构成威胁，致使民众难以接受(Castelo et al., 2019; Ferrari et al., 2016; Yogeeswaran et al., 2016)。人通常自认是独特而不同于其他事物的(Brewer, 1991)，而机器只能按照标准化和模式化的方式运作，以同样的方式处理所有情况(Haslam, 2006)。机器在决策过程中的考量不同于人类，人们因之担心机器决策时忽视自己的独特特征和境遇，比如患者担心人工智难以考虑到其性格和症状的特殊性，从而抵制医疗人工智能(Longoni et al., 2019)。而在日常的个性化推荐场景中，即使算法推荐的准确性胜过人类，人们也不相信算法的推荐能够贴合个人品味(Yeomans et al., 2019)。

算法会将不能被量化的信息删除或简单表征，从而简化信息处理过程，使整个处理过程去情境化(Choi et al., 2007; Nisbett et al., 2001)，但去情境化的决策方式往往忽略人的独特性(Longoni et al., 2019; Sloan & Warner, 2018)。所以在社会化水平较高的任务中，算法因只能分析可量化的指标，而难以辨识品质、态度等指标，而被认为不如人类决策公平和可信(Lee, 2018; Newman et al., 2020)。

拟人化使机器人具备与人相似的外表和相近的能力，构成对人类独特性的威胁，进而影响算法决策趋避 (Yogeeswaran et al., 2016; Złotowski et al., 2017)。倘若区分人类与机器在能力方面的情感相似性，便可以增强人们在主观任务中对算法决策的信任(Castelo et al., 2019)。Hristova 和 Grinberg (2015)在道德困境研究中引入拟人化因素，人们更允许形似人类的机器人而不是人类做出功利主义决策，并给予其更少的道德责备(Hristova & Grinberg, 2015; Hristova & Grinberg, 2016)。有意思的是，同样是选择作为，人类受到的道德责任比机械外观机器人更多，选择不作为时则相反；但人们对类人外观机器人选择作为或

不作为的道德指责模式与对人类的极其相似(Malle et al., 2016)。机器的拟人化程度可能存在一个阈值使得人们对它与人类的道德判断模式几乎无异, 这个阈值具体为何, 仍需探寻。

4　算法决策趋避未尽

试想一下, 若有一人 A 面试一家公司, 发送简历后不久便收到邮件回复。在该邮件中, XC-4110 明确表示自己是人工智能算法 HR, 负责评判 A 的简历并决定是否进入面试环节, 另有 4 位人类 HR 负责其他求职者的简历筛选。在面试阶段, 算法 HR 将与 4 位人类 HR 一起对通过简历筛选的求职者进行面试考核。当 A 了解到负责筛选自己简历的 HR 是人工智能算法, 机器启发式便可能会激活, 进而认为算法比人类更客观、准确、无偏见等, 然后趋近算法决策。但倘若 A 不理解算法如何做出决策或得不到有关决策过程的解释时, 便会回避算法决策。当 A 通过简历筛选进入面试阶段, 便会与算法 HR 面对面交流等。在这一过程中, A 可能认为算法 HR 并非真正人类, 相较于其他 4 位人类 HR 缺乏心智, 难以充分交流, 不具备做出招聘决策的能力, 并怀疑其能否担任相应的后果责任, 从而回避算法 HR 的面试决策。但倘若 A 之前有过被歧视的经历, 他可能认为算法 HR 更一视同仁, 因而趋近算法 HR 决策。当然, 面对算法和人类 HR 时, A 可能会觉得算法 HR 凭借其能力抢占了部分人类工作, 算法与人类有着相似的外表、言语和动作等, 其便会感到现实威胁和身份威胁, 从而回避算法 HR 决策。但当 A 认为算法能力强, 会忽视导致歧视的个人相关信息等时, 便会趋近算法决策。

如上所示, 算法决策趋避的过程动机理论模拟了人们面临算法和人类做出同一决策时的心理动机框架, 并将其嵌入到人机交互的三个阶段中。尽管我们概括现有研究, 提出了该理论框架, 但关于算法决策趋避的研究领域仍有可供讨论之处。

其一, 过程动机理论成立有两个重要条件, 一是算法拟人化; 二是人与算法进行类社会互动。首先, 拟人化能将算法类比于人, 是算法趋近或回避的核心, 若无拟人化, 人们便难以将算法视为社会认知的对象。其次, 在拟人化出现后, 人们可能会根据言语或外貌等非言语特征来判断算法或机器人的性别、年龄和种族(Makatchev et al., 2013; Saunderson & Nejat, 2019), 进行类社会互动并受三种动机作用而对算法产生反应。值得注意的是, 人们会将算法拟人化, 也会对其非人化。研究表明, 无论是人类还是机器人做出伤害行为, 均会降低外界对其心智的感知(Swiderska & Küster, 2020), 未来的研究可以探讨在伤害情境下, 非人化对算法决策趋避的影响。

即使人们无差别地对待人类和算法, 过程动机理论同样适用, 人际交往中尚且有喜好厌恶之分, 在与算法进行类社会互动中也可能如此。就像并非所有白人自由主义者都主张优待弱势群体般, 其中有些人对黑人或少数群体抱有偏见和施以伤害(Bradley-Geist et al., 2010)。在某些决策情境下, 人们可能觉得算法与其他人类并无不同, 但会因自身教育水平(Thurman et al., 2019; van Berkel et al., 2021; Saha et al., 2020)、人工智能技术水平(Schoeffer et al., 2021)而更好地理解或更质疑复杂决策过程, 以及对结果抱有预期和偏好(Jago & Laurin, 2021; Wang et al., 2020)而产生倾向性, 最后表现出算法决策趋避。

其二, 因缺乏研究证据, 在身份认同形成阶段只讨论现实威胁和身份威胁的影响。这可能会导致忽视这样一种情况, 即人们并非总是将算法或机器人视为威胁。有研究发现, 人们也会为机器人举办葬礼(Burch, 2018), 这表明人们可能将机器人视为关系密切的群内成员, 算法决策趋避可能会受此影响。因为人们会对内群体或共同群体表现出内群体偏好(Tajfel & Turner, 1896; Gaertner et al., 1993)。社会认同也包含人类以外的事物(Amiot et al., 2020), 当个体将机器人视为内群体成员后, 会更喜欢与之接触并给予更积极的评价(Eyssel & Kuchenbrandt, 2012)。在人机交互频繁的实现中(Beane, 2019; Newman et al., 2020), 对群内成员的积极态度是否成为趋近算法决策的原因, 对此还需进一步探讨。

其三, 除理论的三个过程阶段外, 应当还考虑是否存在算法决策由回避转为趋近的干预措施。Langer 等(2021)研究表明, 人们最初更信任人类决策而非自动化系统决策, 在出现信任违反并进行修复干预后, 人们对自动化系统决策的信任增量更少。犯错也同样会让人们从最初对无差别的信任转为更加抗拒算法决策(Prahl & van Swol, 2017)。但后续研究可以此开展进行对干预手段的

探讨。

决策本身的内容和适用的情境也可能对算法趋避产生影响。研究表明，人们会在人力资源管理、股市投资、医疗健康和司法等不同决策情境下来表现出算法决策趋避，而这似乎与决策内容或情境所涉及的复杂性、重要性和主观性有关。在教育决策、新闻撰写以及体重估计、吸引力预测等需要进行数据处理、相对客观的场景下(Logg et al., 2019; Marcinkowski et al., 2020)，人们会偏好客观、准确、快速的算法决策(Bonnefon et al., 2016; Donnelly, 2017; Lindebaum et al., 2020)。但在金融投资、战略规划以及道德决策等相对复杂重要需要进行主观考量的情境(Önkal et al., 2009; Leyer & Schneider, 2019; Bigman & Gray, 2018)，去语境化决策的算法决策便会被认为难以胜任(Longoni et al., 2019; Sloan & Warner, 2018; Lee, 2018; Newman et al., 2020)。除此之外，趋近或回避算法决策也可能与自身具体情况有关，即人们从自身利益出发进行考量，从而选择趋避。比如人们认为算法面试推荐更少偏见、歧视和更公平(Jago & Laurin, 2021)，但人们更不信任和拒绝关乎自身安全和利益的算法医疗诊断建议和司法犯罪预测决策等(Bigman & Gray, 2018; Promberger & Baron, 2006)。后续研究还更应探讨决策情境和个性差异对的算法决策趋避的影响。

其四，除理论所概括的三种动机外，人们也可能出于控制等其他动机而选择趋避。目前算法仍可能处于不透明的“黑箱”阶段(Burrell, 2016; Castelvecchi, 2016)，民众难以了解其运行原理(Kroll et al., 2017)，从而感到极大的不确定性(Acikgoz et al., 2020)。不确定和无序引发焦虑并降低控制感，人们在失控后会通过找寻其他途径来补偿或恢复控制感(Kay et al., 2009; Kay & Eibach, 2013)。研究表明，人们厌恶不听人类指令而自主做出决策的机器人(Złotowski et al., 2017)，但若人们能自行调整或控制算法以掌握最终决策权，他们会更愿意接受算法决策(Berger et al., 2021; Dietvorst et al., 2018)。这表明，控制感在一定程度上能逆转人们的算法绝决策回避。然而究竟采用何种策略可以补偿控制感并实现成功逆转是亟需探讨的。同时也要警惕个体长期经历控制感缺失从而最终放弃重拾控制感(Alloy et al., 1984)，这可能会导致人们一直回避算法决策。除了控制动机外，人们也可能会出于成就动机而选择显著优于人类的算法决策，是否如此还需要进一步验证。

其五，本文基于现有实证研究提出自由主义与算法决策趋近有关，而保守主义与算法决策回避有关的观点。但不可否认的是，如同自由主义者也会歧视和伤害黑人和其他少数群体一样(Bradley-Geist et al., 2010)，自由主义者也可能比保守主义者对算法决策带来的隐私风险、现实或身份威胁等问题更为敏感，从而使他们回避算法决策。自由主义并不一定使人们趋近算法决策，可能存在某些因素使自由主义者回避算法决策，保守主义亦然。

最后，在算法决策趋避理论的基础上，应继续完善理论框架，并不断寻找干预措施以遏制其造成的不良影响。张语嫣等(2022)提出的算法拒绝的三维动机理论聚焦于算法拒绝这一现象，归纳了人们会基于算法主体怀疑、道德地位缺失和人类特性湮没这三个主要原因而拒绝算法决策，分别对应信任/怀疑、担责/推责、掌控/失控三种动机。本研究涉及的动机理论则基于广泛的认知、关系和存在动机探讨人们对算法决策的趋避倾向反应，将算法拒绝涵盖于算法决策回避。除了过程动机理论中所提及的原因，算法决策趋避必然还由多种复杂因素而造成，未来还要继续探讨其中因素，并不断完善现有理论或提出新的理论。丹尼尔·卡尼曼等人(2021)的《噪声》一书中指出，人们通过采用决策卫生策略或使用一套既定的算法等进行降噪处理而优化决策，但大众对这些降噪策略或接纳或排斥。人们代之降噪策略的态度似乎与本文中人们对于算法决策或趋近或回避的态度十分相近，但两者却又不同。前者旨在基于原有基础优化决策而减少噪声，只涉及人类或算法单一决策主体；后者则旨在比较人们面对同样的算法决策和人类决策而产生的反应倾向，包含了算法和人类两个决策主体。两者虽略不同，但倘若比较人们对降噪处理方式完全一致的人类和算法决策的反应，便可得出人们的算法决策趋避倾向。因此，未来研究可以噪声为切入点来进一步探究算法决策趋避，借鉴降噪策略中的深刻见解、处理原则等来干预、优化算法决策，使之更

好服务人类, 增进人类福祉。

参考文献

丹尼尔 • 卡尼曼, 奥利维耶·西博尼, 卡斯·R. 桑斯坦. (2021). *噪声* (李纾, 汪祚军, 魏子晗, 译). 浙江教育出版社.

许丽颖, 喻丰, 彭凯平. (2022). 算法歧视比人类歧视引起更少道德惩罚欲. *心理学报, 54*(9), 1076–1092.

许丽颖, 喻丰, 邬家骅, 韩婷婷, 赵靓. (2017). 拟人化: 从 "它" 到 "他". *心理科学进展*, *25*(11), 1942–1954.

喻丰, 许丽颖. (2020). 人工智能之拟人化. *西北师大学报: 社会科学版*, *57*(5), 52–60.

张语嫣, 许丽颖, 喻丰, 丁晓军, 邬家骅, 赵靓. (2022). 算法拒绝的三维动机理论. *心理科学进展*, *30*(5), 1093–1105.

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, *28*(4), 399–416.

Alloy, L. B., Peterson, C., Abramson, L. Y., & Seligman, M. E. (1984). Attributional style and the generality of learned helplessness. *Journal of Personality and Social Psychology*, *46*(3), 681–687.

Amiot, C. E., Sukhanova, K., & Bastian, B. (2020). Social identification with animals: Unpacking our psychological connection with other animals. *Journal of Personality and Social Psychology*, *118*(5), 991–1017.

Anderson, C. J., & Singer, M. M. (2008). The sensitive left and the impervious right: Multilevel models and the politics of inequality, ideology, and legitimacy in Europe. *Comparative Political Studies*, *41*(4–5), 564–599.

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623.

Auxier, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). Americans and privacy: Concerned, confused, and feeling lack of control over their personal information. *Pew Research Center: Internet, Science & Tech.* Retrieved January 15, 2022, from https://policycommons.net/artifacts/616499/americans-and-privacy/1597152/on 13 Jun 2022. CID: 20.500.12592/hx524v

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*(2), 134–143.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., ... de Souza, A. F. (2020). Self-driving cars: A survey. *Expert Systems with Applications*, *165*, Article 113816. https://doi.org/10.1016/j.eswa.2020.113816

Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditional automatic attitude activation with a pronunciation task. *Journal of Personality and Social Psychology*, *32*(1), 104–128.

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. (2012). Don't mind meat? The denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin*, *38*(2), 247–256.

Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, *64*(1), 87–123.

Benzell, S. G., Kotlikoff, L. J., LaGarda, G., & Sachs, J. D. (2015). *Robots are us: Some economics of human replacement* (No. w20941). National Bureau of Economic Research. Cambridge, MA.

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, *63*(1), 55–68.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34.

Bigman, Y. E., Gray, K., Waytz, A., Arnestad, M., & Wilson, D. (2020). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Expreimental Psychology: General*, Advance online publication. https://doi.org/10.31234/osf.io/m3nrp

Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J., & Gray, K. (2021). Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior, 122*, Article 106859. https://doi.org/10.1016/j.chb.2021.106859

Biró, P., van de Klundert, J., Manlove, D., Pettersson, W., Andersson, T., Burnapp, L., … Viana, A. (2021). Modelling and optimisation in European kidney exchange programmes. *European Journal of Operational Research*, *291*(2), 447–456.

Blanchard, D. C., Hynd, A. L., Minke, K. A., Minemoto, T., & Blanchard, R. J. (2001). Human defensive behaviors to threat scenarios show parallels to fear-and anxiety-related defense patterns of non-human mammals. *Neuroscience & Biobehavioral Reviews*, *25*(7–8), 761–770.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151.

Bonanno, G. A., & Jost, J. T. (2006). Conservative shift among high-exposure survivors of the September 11th terrorist attacks. *Basic and Applied Social Psychology, 28*, 311–323.

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447–459.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Bostrom, N. (2014). *Superintelligence*. New York: Oxford University Press.

Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., & McKenna, C. (2010). Moral credentialing by association: The importance of choice and relationship closeness. *Personality and Social Psychology Bulletin*, *36*(11), 1564–1575.

Brandt, M. J., & Reyna, C. (2011). The chain of being: A hierarchy of morality. *Perspectives on Psychological Science*, *6*(5), 428–446.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin*, *17*(5), 475–482.

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development*, *90*(4), 1202–1214.

Bryson, J. J. (2020). The artificial intelligence of the ethics of artificial intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Eds.). *The oxford handbook of ethics of AI* (pp. 3–25). New York: Oxford University press.

Burch, J. (2018). *AIBO robots dogs given buddhist funeral in Japan*. Retrieved February 4, 2022, from https://www.nationalgeographic.com/travel/article/in-japan--a-buddhist-funeral-service-for-robot-dogs

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), Article 2053951715622512. https://doi.org/10.1177/2053951715622512

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239.

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825.

Castelo, N., & Ward, A. F. (2021). Conservatism predicts aversion to consequential artificial intelligence. *Plos One*, *16*(12), Article e0261467. https://doi.org/10.1371/journal.pone.0261467

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature, 538*(7623), 20–23.

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1), Article 100698. https://doi.org/10.1016/j.hrmr.2019.100698

Choi, I., Koo, M., & Choi, J. A. (2007). Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin*, *33*(5), 691–705.

Chung, S., & Cho, H. (2017). Fostering parasocial relationships with celebrities on social media: Implications for celebrity endorsement. *Psychology & Marketing, 34*(4), 481–495.

Cooley, E., Brown-Iannuzzi, J., & Cottrell, D. (2019). Liberals perceive more racism than conservatives when police shoot Black men—But, reading about White privilege increases perceived racism, and shifts attributions of guilt, regardless of political ideology. *Journal of Experimental Social Psychology*, *85*, Article 103885. https://doi.org/10.1016/j.jesp.2019.103885

Cooley, E., & Payne, B. K. (2017). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin*, *43*(1), 46–59.

Cooley, E., & Payne, B. K. (2019). A group is more than the average of its parts: Why existing stereotypes are applied more to the same individuals when viewed in groups than when viewed alone. *Group Processes & Intergroup Relations*, *22*(5), 673–687.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Davis, M. (1973). *Intimate Relations*. New York, NY: Free Press.

Diab, D. L., Pui, S. Y., Yankelevich, M., & Highhouse, S. (2011). Lay perceptions of selection decision aids in US and non-US samples. *International Journal of Selection and Assessment*, *19*(2), 209–216.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170.

Dixon, G., Hart, P. S., Clarke, C., O'Donnell, N. H., & Hmielowski, J. (2020). What drives support for self-driving car technology in the United States?. *Journal of Risk Research*, *23*(3), 275–287.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019, March). Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 275–285). California.

Donnelly, L. (2017). Forget your GP, robots will 'soon be

able to diagnose more accurately than almost any doctor'. *The Telegraph*. Retrieved October 21, 2021, from https://www.telegraph.co.uk/technology/2017/03/07/robots–will–soon–able–diagnose–accurately–almost–doctor/

Ehsan, U., & Riedl, M. (2019). On design and evaluation of human-centered explainable AI systems. *Glasgow '19*. Scotland, American.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, *34*(3), 169–189.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542* (7639), 115–118.

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, *51*(4), 724–731.

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring human-machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics*, *8*(2), 287–302.

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2021). Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, *31*, Article 100524. https://doi.org/10.1016/j.jbef.2021.100524

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. New York: Mcgraw-Hill Book Company.

Gaertner, S. L., Dovidio, J. F., Anastasio, P. A., Bachman, B. A., & Rust, M. C. (1993). The common ingroup identity model: Recategorization and the reduction of intergroup bias. *European Review of Social Psychology*, *4*(1), 1–26.

Gary, M. S., & Wood, R. E. (2011). Mental models, decision rules, and performance heterogeneity. *Strategic Management Journal*, *32*(6), 569–594.

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review*, *77*(2), 167–187.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge: Cambridge university press.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*, 627–660.

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*, 97–103.

Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610.

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–25.

Grgić-Hlača, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018 April). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference* (pp. 903–912). Geneva.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018 February). Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1). New Orleans, Louisiana.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., ... Zalaudek, I. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842.

Han, S., & Yang, H. (2018). Understanding adoption of intelligent personal assistants: A parasocial relationship perspective. *Industrial Management & Data Systems, 118*(3), 618–636.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527.

Hartmann, T. (2008). Parasocial interactions and paracommunication with new media characters. In E. A. Konijn, S. Utz, M. Tanis, & S. B. Barnes (Eds.), *Mediated Interpersonal Communication* (pp. 177–199). New York, NY: Routledge.

Harvey, C. R., Rattray, S., Sinclair, A., & van Hemert, O.

(2017). Man vs. machine: Comparing discretionary and systematic hedge fund performance. *The Journal of Portfolio Management*, *43*(4), 55–69.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264.

Headrick, D. R. (2009). *Technology: A world history*. New York: Oxford University Press.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science*, *9*(4), 393–401.

Helberger, N., Araujo, T., & de Vreese, C. H. (2020). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review*, *39*, Article 105456. https://doi.org/10.1016/j.clsr.2020.105456

Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, *19*(3), 215–229.

Howard, F. M., Gao, C. A., & Sankey, C. (2020). Implementation of an automated scheduling tool improves schedule quality and resident satisfaction. *Plos One*, *15*(8), Article e0236952. https://doi.org/10.1371/journal.pone.0236952

Hristova, E., & Grinberg, M. (2015). Should robots kill? Moral judgments for actions of artificial cognitive agents. In *Proceedings of Euro Asian Pacific Joint Conference on Cognitive Science* (pp. 306–311). Torino, Italy.

Hristova, E., & Grinberg, M. (2016). Should moral decisions be different for human and artificial cognitive agents. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1511–1516). Austin, TX.

Hu, P. J. H., Hu, H. F., & Fang, X. (2017). Examining the mediating roles of cognitive load and performance outcomes in user satisfaction with a website: A field quasi-experiment. *MIS Quarterly*, *41*(3), 975–987.

Huang, H. L., Cheng, L. K., Sun, P. C., & Chou, S. J. (2021). The effects of perceived identity threat and realistic threat on the negative attitudes and usage intentions toward hotel service robots: The moderating effect of the robot's anthropomorphism. *International Journal of Social Robotics*, *13*(7), 1599–1611.

Iacoviello, V., & Spears, R. (2021). Playing to the gallery: Investigating the normative explanation of ingroup favoritism by testing the impact of imagined audience. *Self and Identity*. Advance online publication. https://doi.org/10.1080/15298868.2021.1933582

Ireland, L. (2020). Who errs? Algorithm aversion, the source of judicial error, and public support for self-help behaviors. *Journal of Crime and Justice*, *43*(2), 174–192.

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1), 38–56.

Jago, A. S., & Laurin, K. (2021). Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin*, *48*(4), 582–595.

Jia, C., & Liu, R. (2021). Algorithmic or human source? Examining relative hostile media effect with a transformer-based framework. *Media and Communication*, *9*(4), 170–181.

Jost, J. T., & Amodio, D. M. (2012). Political ideology as motivated social cognition: Behavioral and neuroscientific evidence. *Motivation and Emotion*, *36*(1), 55–64.

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology*, *60*, 307–337.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375.

Jost, J. T., Ledgerwood, A., & Hardin, C. D. (2008a). Shared reality, system justification, and the relational basis of ideological beliefs. *Social and Personality Psychology Compass*, *2*(1), 171–186.

Jost, J. T., Nosek, B. A., & Gosling, S. D. (2008b). Ideology: Its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science*, *3*(2), 126–136.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, *93*(5), 1449–1475.

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019). Applicant perceptions of hiring algorithms-uniqueness and discrimination experiences as moderators. In *Academy of Management Proceedings* (Vol. 2019, No. 1, p. 18172). Briarcliff Manor, NY 10510: Academy of Management.

Kay, A. C., & Eibach, R. P. (2013). Compensatory control and its implications for ideological extremism. *Journal of Social Issues*, *69*(3), 564–585.

Kay, A. C., Whitson, J. A., Gaucher, D., & Galinsky, A. D. (2009). Compensatory control: Achieving order through the mind, our institutions, and the heavens. *Current Directions in Psychological Science, 18*(5), 264–268.

Komatsu, T. (2016, March). Japanese students apply same moral norms to humans and robot agents: Considering a moral HRI in terms of different cultural and academic backgrounds. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 457–458). Christchurch, New Zealand.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. M. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*,

8(1), 1–11.

Kramer, M. F., Schaich Borg, J., Conitzer, V., & Sinnott-Armstrong, W. (2018, December). When do people want AI to make decisions?. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 204–209). New Orleans, Louisiana.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633–704.

Krupp, M. M., Rueben, M., Grimm, C. M., & Smart, W. D. (2017, March). Privacy and telepresence robotics: What do non-scientists think?. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 175–176). Vienna, Austria.

Kugler, M., Jost, J. T., & Noorbaloochi, S. (2014). Another look at moral foundations theory: Do authoritarianism and social dominance orientation explain liberal-conservative differences in "moral" intuitions?. *Social Justice Research*, *27*(4), 413–431.

Langer, M., König, C. J., Back, C., & Hemsing, V. (2021). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *PsyArXiv Prepints*. Advance online publication. https://doi.org/10.31234/osf.io/r9y3t

Langer, M., König, C. J., & Fitili, A. (2018). Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection. *Computers in Human Behavior, 81*, 19–30.

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217–234.

Lasota, P. A., Fong, T., & Shah, J. A. (2017). A survey of methods for safe human-robot interaction. *Foundations and Trends® in Robotics*, *5*(4), 261–349.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1–16.

Lee, M. K., & Baykal, S. (2017, February). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 Acm Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1035–1048). New York, USA.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26.

Leyer, M., & Schneider, S. (2019, June). Me, you, or AI? How do we feel about delegation. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Stockholm & Uppsala, Sweden.

Li, J., Zhao, X., Cho, M. J., Ju, W., & Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *SAE Technical Paper*, *10,* 1–8.

Lima, G., Kim, C., Ryu, S., Jeon, C., & Cha, M. (2020). Collecting the public perception of AI and robot rights. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–24.

Lindebaum, D., Vesa, M., & den Hond, F. (2020). Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review, 45*(1), 247–263.

Liu, B., & Wei, L. (2019). Machine authorship in situ: Effect of news organization and news genre on news credibility. *Digital Journalism*, *7*(5), 635–657.

Logg, J. M. (2017). Theory of machine: When do people rely on algorithms?. *Harvard Business School working paper.* 17–86. Advance online publication. https://dash.harvard.edu/handle/1/31677474

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, *49*(5), 773–785.

Mahadevan, K., Somanath, S., & Sharlin, E. (2018, March). "Fight-or-flight" leveraging instinctive human defensive behaviors for safe human-robot interaction. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 183–184). Chicago, USA.

Makatchev, M., Simmons, R., Sakr, M., & Ziadee, M. (2013, March). Expressing ethnicity through behaviors of a robot character. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 357–364). Tokyo, Japan.

Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243-256.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., &

Cusimano, C. (2015, March). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 117–124). New York, American.

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016, March). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 125–132). Christchurch, New Zealand.

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020, January). Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice, and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 122–130). Barcelona, Spain.

Martínez, E., & Winter, C. (2021). Protecting sentient artificial intelligence: A survey of lay intuitions on standing, personhood, and general legal protection. *Frontiers in Robotics and AI*, *8*, Article 788355. https://doi.org/10.3389/frobt.2021.788355

McClure, P. K. (2018). “You’re fired,” says the robot: The rise of automation in the workplace, technophobes, and fears of unemployment. *Social Science Computer Review*, *36*(2), 139–156.

McFarland, M. (2014). Elon Musk: ‘With artificial intelligence we are summoning the demon’. *The Washington Post*. Retrieved June 19, 2021, from https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon–musk–with–artificial–intelligence–we–are–summoning–the–demon/?utm_term=.02d648908751

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*(3), 370–375.

Miller, S. M., & Keiser, L. R. (2021). Representative bureaucracy and attitudes toward automated decision making. *Journal of Public Administration Research and Theory*, *31*(1), 150–165.

Möhlmann, M., & Zalmanson, L. (2017, December). Hands on the wheel: Navigating algorithmic management and Uber drivers’. In *Autonomy’, in Proceedings of the International Conference on Information Systems (ICIS)* (pp.10–13). Seoul, Korea.

Napier, J. L., & Jost, J. T. (2008). Why are conservatives happier than liberals?. *Psychological Science*, *19*(6), 565–572.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates?. *International Journal of Human-Computer Studies*, *45*(6), 669–678.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender‐stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, *27*(10), 864–876.

Neal-Barnett, A., Stadulis, R., Singer, N., Murray, M., & Demmings, J. (2010). Assessing the effects of experiencing the acting White accusation. *The Urban Review*, *42*(2), 102–122.

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*(2), 291–310.

Niszczota, P., & Kaszás, D. (2020). Robo-investment aversion. *Plos One*, *15*(9), Article e0239277. https://doi.org/10.1371/journal.pone.0239277

Noble, S. M., Foster, L. L., & Craig, S. B. (2021). The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment*. Advance online publication. https://doi.org/10.1111/ijsa.12320

Nolan, K. P., Carter, N. T., & Dalal, D. K. (2016). Threat of technological unemployment: Are hiring managers discounted for using standardized employee selection practices?. *Personnel Assessment and Decisions*, *2*(1), 34–47.

Nolan, K. P., Dalal, D. K., & Carter, N. (2020). Threat of technological unemployment, use intentions, and the promotion of structured interviews in personnel selection. *Personnel Assessment and Decisions*, *6*(2), 38–53.

Nørskov, S., Damholdt, M. F., Ulhøi, J. P., Jensen, M. B., Ess, C., & Seibt, J. (2020). Applicant fairness perceptions of a robot-mediated job interview: A video vignette-based experimental survey. *Frontiers in Robotics and AI*, *163*, Article 586263. https://doi.org/10.3389/frobt.2020.586263

Oleksy, T., & Wnuk, A. (2021). Do women perceive sex robots as threatening? The role of political views and presenting the robot as a female-vs male-friendly product. *Computers in Human Behavior*, *117*, Article 106664. https://doi.org/10.1016/j.chb.2020.106664

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409.

Ötting, S. K., & Maier, G. W. (2018). The importance of

procedural justice in human-machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior*, *89*, 27–39.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Peng, Y. (2020). The ideological divide in public perceptions of self-driving cars. *Public Understanding of Science*, *29*(4), 436–451.

Pezzo, M. V., & Pezzo, S. P. (2006). Physician evaluation after medical errors: Does having a computer decision aid help or hurt in hindsight? *Medical Decision Making, 26*(1), 48–56.

Prahl, A., & van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted?. *Journal of Forecasting*, *36*(6), 691–702.

Promberger, M., & Baron, J. (2006). Do patients trust computers?. *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Radinsky, W. (2015). Robotics, AI, the luddite fallacy and the future of the job market. In B. Goertzel & T. Goertzel (Eds.), *The end of the beginning: Life, society and economy on the brink of the singularity* (pp. 159–185). Los Angeles, CA: Humanity+ Press.

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, *46*(1), 192–210.

Reinecke, M. G., Wilks, M., & Bloom, P. (2021, July). Developmental changes in perceived moral standing of robots. In *Proceedings of the 43nd Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43). Vienna, Austria.

Riek, B. M., Mania, E. W., & Gaertner, S. L. (2006). Intergroup threat and outgroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *10*(4), 336–353.

Rubin, R. B., & Rubin, A. M. (2001). Attribution in social and parasocial relationships. In M. A. Fitzpatrick, H. Reis, & A. Vangelista (Eds.), *Attribution, communication behavior, and close relationships* (pp. 320–337). Cambridge: Cambridge University Press.

Saha, D., Schumann, C., Mcelfresh, D., Dickerson, J., Mazurek, M., & Tschantz, M. (2020, November). Measuring non-expert comprehension of machine learning fairness metrics. In *Proceedings of the 37th International Conference on Machine Learning* (8377–8387). *PMLR*.

Saunderson, S., & Nejat, G. (2019). How robots influence humans: A survey of nonverbal communication in social human-robot interaction. *International Journal of Social Robotics*, *11*(4), 575–608.

Scheutz, M., & Malle, B. F. (2021). May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). In J. Galliot, D. MacInosh, & J. D. Ohlin (Eds.), *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare* (pp. 89–102). New York: Oxford University Press.

Schlicker, N., Langer, M., Ötting, S., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up ‘Black Boxes’? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*, *122*, Article 106837. https://doi.org/10.1016/j.chb.2021.106837

Schneirla, T. C. (1959). An evolutionary and developmental theory of biphasic processes underlying approach and withdrawal. In M. R. Jones (Ed.), Nebraska symposium on motivation (pp. 1–42). Univer: Nebraska Press.

Schoeffer, J., Machowski, Y., & Kuehl, N. (2021). Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making. *arXiv preprint arXiv:2109.05792*.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565.

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*, 277–284.

Silva, S., & Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon*, *55*(1 & 2), 9–37.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint* arXiv:1712.01815.

Sloan, R. H., & Warner, R. (2018). When is an algorithm transparent? Predictive analytics, privacy, and public policy. *IEEE Security & Privacy*, *16*(3), 18–25.

Steffel, M., & Williams, E. F. (2018). Delegating decisions: Recruiting others to make choices we might regret. *Journal of Consumer Research*, *44*(5), 1015–1032.

Steffel, M., Williams, E. F., & Perrmann-Graham, J. (2016). Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes*, *135*, 32–44.

Stern, B. B., Russell, C. A., & Russell, D. W. (2007). Hidden persuasions in soap operas: Damaged heroines and negative consumer effects. *International Journal of Advertising*, *26*(1), 9–36.

Stewart, B. D., & Morris, D. S. (2021). Moving morality beyond the in-group: Liberals and conservatives show

differences on group-framed moral foundations and these differences mediate the relationships to perceived bias and threat. *Frontiers in Psychology*, *12,* Article 579908. doi: 10.3389/fpsyg.2021.579908.

Stewart, B. D., Gulzaib, F., & Morris, D. S. (2019). Bridging political divides: Perceived threat and uncertainty avoidance help explain the relationship between political ideology and immigrant attitudes within diverse intergroup contexts. *Frontiers in Psychology*, *10*, 1236–1254.

Suen, H. Y., Chen, M. Y. C., & Lu, S. H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes?. *Computers in Human Behavior*, *98*, 93–101.

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & Flanagin, A. J (Eds.), *Digital media, youth, and credibility* (pp. 73–100). Cambridge, MA: The MIT Press.

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–9). Glasgow, Scotland.

Sundar, S. S., Kim, J., Rosson, M. B., & Molina, M. D. (2020, April). Online privacy heuristics that predict information disclosure. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Honolulu.

Suresh, A., Latha, S. S., Nair, P., & Radhika, N. (2014). Prediction of fight or flight response using artificial neural networks. *American Journal of Applied Sciences*, *11*(6), 912–920.

Swiderska, A., & Küster, D. (2020). Robots as malevolent moral agents: Harmful behavior results in dehumanization, not anthropomorphism. *Cognitive Science*, *44*(7), Article e12872. https://doi.org/10.1111/cogs.12872

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social science information*, *13*(2), 65–93.

Tajfel, H. (Ed.). (2010). *Social identity and intergroup relations* (Vol. 7). Cambridge: Cambridge University Press.

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of intergroup relations* (pp. 7–24). Chicago, IL: Nelson-Hall.

Tandoc, E. C., Jr., Yao, L. J., & Wu, S. (2020). Man vs. machine? The impact of algorithm authorship on news credibility. *Digital Journalism*, *8*(4), 548–562.

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism*, *7*(4), 447–469.

Todd, P., & Benbasat, I. (1994). The influence of decision aids on choice strategies: An experimental analysis of the role of cognitive effort. *Organizational Behavior and Human Decision Processes*, *60*(1), 36–74.

Tran, G. A., Yazdanparast, A., & Strutton, D. (2019). Investigating the marketing impact of consumers' connectedness to celebrity endorsers. *Psychology & Marketing*, *36*(10), 923–935.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Oxford, UK: Basil Blackwell.

Turner, R. N., West, K., & Christie, Z. (2013). Out‐group trust, intergroup anxiety, and out‐group attitude as mediators of the effect of imagined intergroup contact on intergroup behavioral tendencies. *Journal of Applied Social Psychology*, *43*(S2), E196–E205.

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021, May). Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). Yokohama, Japan.

van der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology*, *54*, 93–100.

van Vugt, M., & Hart, C. M. (2004). Social identity as social glue: The origins of group loyalty. *Journal of Personality and Social Psychology*, *86*(4), 585–598.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016, August). Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780). Christchurch, New Zarland.

Waddell, T. F. (2019). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100.

Wang, A. J. (2018). Procedural justice and risk-assessment algorithms. *SSRN Electronic Journal*. Article 3170136. http://dx.doi.org/10.2139/ssrn.3170136

Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Honolulu.

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital*

Journalism, *9*(1), 64–83.
Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117.
Wetherell, G. A., Brandt, M. J., & Reyna, C. (2013). Discrimination across the ideological divide: The role of value violations and abstract values in discrimination by liberals and conservatives. *Social Psychological and Personality Science*, *4*(6), 658–667.
Whang, C., & Im, H. (2021). "I like your suggestion!" The role of humanlikeness and parasocial relationship on the website versus voice shopper's perception of recommendations. *Psychology & Marketing*, *38*(4), 581–595.
Wilson, G. (1973). *The psychology of conservatism* (p. 277). Oxford, England: Academic Press.
Wölker, A., & Powell, T. E. (2021). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism*, *22*(1), 86–103.
Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414.
Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human-Robot Interaction*, *5*(2), 29–47.
Zajonc, R. (1998). Emotion. In Gilbert, D. T., Fiske, S. T., & Lindzey, G. (Eds.). *The handbook of social psychology* (Vol. 1, pp. 591–632). New York: Oxford University Press.
Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies*, *100*, 48–54.

The process motivation model of algorithmic decision-making approach and avoidance

XIE Caifeng[1], WU Jiahua[1], XU Liying[2], YU Feng[1], ZHAND Yuyan[1], XIE Yingying[3]
([1] *Department of Psychology, School of Philosophy, Wuhan University, Wuhan* 430072, *China*)
([2] *School of Social Marxism, Tsinghua University, Beijing* 100084, *China*)
([3] *Lingshan County, Nalong Town, Central School, Qinzhou* 535414, *China*)

Abstract: Algorithms are often used for decision-making. However, algorithmic decision-making is more related to different responses in individuals than human decision-making on the same content. The phenomenon is defined as the algorithmic decision-making approach and avoidance. The approach means that algorithmic decision-making is considered fairer, less biased, less discriminatory, more trustworthy, and more acceptable than human decision-making. But the avoidance is the other way around. To explain the phenomenon of the algorithmic decision-making approach and avoidance better, the process motivation model of algorithmic decision-making approach and avoidance is employed in the review. It summarizes three stages of the interaction between human and algorithm, namely, the interaction of initial behavior, the establishment of quasi-social relationship and the formation of identity. Moreover, it elaborates how cognitional, relational, and existential motivation trigger individual approach and avoidance responses in each specific stage. For future directions, we suggest that more researches are needed to explore how mind perception and intergroup perception influence algorithmic decision-making approach and avoidance. Meanwhile, what is the reversal process of algorithmic decision-making approach and avoidance from a more social perspective and what other possible motivations are associated with it are also worth of considered.

Keywords: algorithmic decision-making, human decision-making, decision-making approach and avoidance, mental motivation, human-algorithm interaction